DOCUMENT RESUME

ED 298 173 TM 012 322

AUTHOR McLarty, Joyce R.
TITLE A Single-Minded Appraisal of a Multiple-Criterion Teacher Appraisal System.
PUB DATE Apr 87
NOTE 23p.; Paper presented at the Annual Meeting of the American Educational Research Association (Washington, DC, April 20-24, 1987). Appendix contains small print.
PUB TYPE Speeches/Conference Papers (150) -- Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Career Ladders; Elementary Secondary Education; Evaluation Methods; Personnel Evaluation; Professional Development; *Program Evaluation; State Programs; *Teacher Evaluation; Teacher Qualifications; Testing Programs
IDENTIFIERS Multidimensional Approach; Teacher Appraisal Instrument; *Tennessee Career Ladder Program

ABSTRACT

The Tennessee Career Ladder evaluation system is a multiple-data-source system for evaluating teachers in order to provide them with merit status and benefits. This presentation is based primarily on data from 1983 through Spring 1986, and focus is on the functioning of the general education teacher appraisal system. Four of the 14 steps involved in designing a career ladder evaluation system are emphasized: (1) formulate criteria/competencies to be assessed; (2) develop data collection instruments/procedures; (3) establish the scoring system; and (4) set standards and cutoff scores. Advantages and disadvantages of each are discussed. Currently, the program utilizes two hours of classroom observation on three occasions; three hour-long dialogues; a professional development and leadership summary; student questionnaires; a principal questionnaire; written tests of reading, writing, and professional knowledge; and an evaluator consensus judgment. The major disadvantage of the system is its level of sophistication. The complexity of good teaching, the high-stakes nature of the award (up to $35,000 per candidate over a five-year period), and the lack of any quota system (no external controls on the total cost of the awards) led the developers of the program to provide the most accurate, fair, and carefully equated program possible. An appendix contains data sources for the competencies and indicators evaluated. (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED298173

A Single-Minded Appraisal of a Multiple-Criterion Teacher Appraisal System

Symposium Presentation By

Joyce R. McLarty, Ph.D.

Director
Test Development Department
Assessment Programs Area
Test Development Division
ACT

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Joyce R. McLarty

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

American Educational Research Association
Annual Meeting
Washington, D. C.
April, 1987

TM 012 322

BEST COPY AVAILABLE

Abstract

The Tennessee Career Ladder evaluation system is a multiple-data-source system for evaluating teachers in order to provide them with merit status and benefits. Currently, the program utilizes two hours of classroom observation on each of three occasions; three hour-long dialogues; a professional development and leadership summary; student questionnaires; a principal questionnaire; written tests of reading, writing and professional knowledge; and an evaluator consensus judgment.

Since its implementation almost three years ago, the program has been strengthened and improved. Most of the original problems have been solved, and those remaining appear to be solvable. Although a definitive validation study has yet to be completed, the evaluation system appears to be identifying meritorious teachers accurately in most cases.

The major disadvantage of the system, as it has been developed, is its level of sophistication. The complexity of good teaching, the high-stakes nature of the award (up to $35,000 per candidate over a five year period), and the lack of any quota system (no external controls on the total cost of the awards), led the developers of the program to provide the most accurate, fair, and carefully equated program possible. Such a program is necessarily costly and complex. However, the additional accuracy and fairness possible with the multiple-data-source approach is essential where Career Ladder decisions which have substantial financial and personal impact must be made.

Overview

In 1983, the Tennessee State Department of Education began a program to identify and reward Tennessee's best teachers. This Career Ladder program was designed primarily as a merit pay system. That is, teachers are to be rewarded for being outstanding teachers, not for performing extra duties. For those not familiar with the Tennessee Career Ladder program, a brief summary may be in order. This program differs from most others in three ways: (1) it is funded and conducted by the state, (2) participation above the lowest level is optional, and (3) it utilizes multiple data sources (currently, two hours of classroom observation on each of three occasions; three hour-long dialogues; a professional development and leadership summary; student questionnaires; a principal questionnaire; written tests of reading, writing and professional knowledge; and an evaluator consensus judgment). More detailed information on these data sources is available in McLarty (1987a), in Furtwengler, Malo, McLarty and Strouss (1986), and in McLarty, Furtwengler, and Malo (1985).

Evaluations are conducted by three highly-trained peer teachers from a system other than the candidate's, at least one of whom is experienced in the same grade level and subject area as the candidate. The peer teachers are given leave from their school systems and serve one to two years full time as evaluators for the state.

This is the third year of implementation of the Tennessee Career Ladder program. Developed in 1983-84, the program experienced its first, and most tempestuous, year in 1984-85. After that year, a number of changes in instrumentation and procedures were made. These survived 1985-86 quite well, and the 1986-87 evaluation system is very similar to the 1986-85 one (the most notable change being from a year-long to a semester-long evaluation cycle).

This presentation is based primarily on information from the 1985-86 year and before, as data from first semester evaluations conducted in fall, 1986, are not yet analyzed. Although most of the remarks are directed toward the functioning of the general education teacher appraisal system, which was the earliest implemented and continues to have the largest numbers of candidates, many other systems are in place and some are now in their second year of implementation. There are systems to evaluate vocational education, chapter 1, and special education (two separate systems depending on the functional level of the students) teachers; librarians; elementary and secondary counselors; principals and assistant principals; and instructional supervisors. Additional systems in development include ones

for speech and language teachers and school psychologists. Eventually, it is expected that systems will be available for all certificated personnel.

Fourteen steps have been identified as necessary in the design of a Career Ladder evaluation system (see Figure 1). However, this presentation will focus on four of them: (1) formulate criteria/competencies to be assessed, (4) develop instruments/procedures for data collection, (6) establish the scoring system, and (11) set standards and cutoff scores. Two additional areas, the nature of assumptions underlying the Career Ladder evaluations and the validation of the Career Ladder evaluation system will be discussed (assumptions first, validation last). In the discussion below, each step, as it was taken by the Tennessee Career Ladder evaluation system, is described and the major advantages and disadvantages of the approach taken are noted.

Assumptions underlying the Tennessee Career Ladder Evaluation System

The assumptions underlying Career Ladder assessments seem to receive somewhat less attention than the political circumstances surrounding Career Ladder programs, the competencies assessed, the instruments used, the scoring procedures, the results, the costs, or even the candidates' opinions of the evaluators. However, there is no more important aspect of Career Ladder evaluation than its assumptions. There is always, of course, the assumption that monetary rewards provided to meritorious teachers will somehow improve the education profession, perhaps by retaining the best teachers in the classroom, by attracting highly competent individuals into the field of education, or by providing a source of motivation and information which will help current teachers to improve their teaching.

The effort to improve education indirectly (by recruiting, retaining, and rewarding the best teachers) was the focus and assumption of the Tennessee Career Ladder program from its initiation. However, during the program's tenure, a large percentage of the modifications in it have been designed to increase the program's ability to directly improve education by providing teachers with diagnostic information and multiple opportunities to improve their performance sufficiently to qualify. Because the original system was designed as a summative rather than formative evaluation, some of these changes have actually undermined the validity of the decision process.

In addition to the assumptions underlying the creation of the Career Ladder program, a number of additional assumptions were necessary to its implementation. Most

important was the assumption that meritorious teaching was a characteristic of the teacher. If teachers were to be considered meritorious across all of the subject areas and grade levels and student types for which they were endorsed, then teaching excellence had to be assumed to generalize across those context conditions. Without this assumption, it would not have been appropriate to continue to provide a merit award for a teacher who had been evaluated teaching social studies to students in grade seven when that teacher went on to teach language arts to students in grade eight.

Another assumption made for the Tennessee Career Ladder program was that all teachers basically performed the same activities; some just performed them more consistently or better than others. This assumption led, eventually, to a single evaluation system with multiple cutoff scores (standards). Had this assumption not been made, it would have been appropriate to design a separate, qualitatively distinct evaluation system for each level of the Career Ladder.

Finally, it was assumed that a number of teaching styles were effective and appropriate. No single style (for example, direct instruction) was considered to be the way to teach for all teachers under all circumstances. It would have been considerably easier to design an evaluation system to reward a single style of teaching (and easier as well for the teachers to understand what was required under that system), but the assumption was made that multiple styles were appropriate and a single style should not be dictated.

It should be noted that the lack of two features in the Tennessee Career Ladder evaluation system is not based on assumptions. First, the lack of outcome measures of student performance is not because these are not considered important aspects to assess in determining a teacher's merit, but rather because the procedures and instruments for measuring them appropriately are still being pilot tested. Second, the lack of an assessment of subject area competence is not because it is not considered important, but because appropriate instruments were not available for use in measuring it. Consideration is being given to including it in the future.

Formulate criteria/competencies to be assessed

The first step in developing the Career Ladder evaluation system was to define the objectives to be measured. These, together with the measurement methodology, would constitute the working definition of what Tennessee means when it identifies as teacher as meritorious. First a

number of possible characteristics were selected from the research literature. These were reviewed, added to, and amended by Tennessee's teachers. Indeed, the first efforts resulted in hundreds of possible areas (objectives) for evaluation of teachers.

Once this large set of objectives was assembled, it was reduced in three ways. First, those objectives which could not be reasonably measured in a statewide evaluation system were set aside for future consideration. Then objectives which were appropriate to only a few teachers or circumstances were eliminated. Finally, the remaining objectives were organized into a few, broad categories. The resulting domains of measurement were: Planning, Teaching Strategies, Evaluation, Classroom Management, and Leadership. These, and their subcategories (indicators of measurement) are provided in Appendix A.

Advantages

The most important advantages of this approach to identifying content were the use of the research base and the involvement of Tennessee teachers. The research provided an excellent starting point in defining characteristics of meritorious teachers. It also helped to establish the credibility of the system.

The involvement of Tennessee teachers similarly helped to establish credibility, but moreover it provided the test of reasonability for the content. Tennessee's teachers broadened the concepts from the research (many of which were expressed in such narrow terms that they became trivial), and extended them beyond the methodology of direct instruction. In addition, they raised real and profound questions about what, exactly, a teacher should be held accountable for. Many of the tentatively considered objectives which would have held the teacher accountable for, for example, locating and using appropriate media supplements to the instruction, ended up "on the cutting room floor" after discussions relating to the teacher's role and resources in actual school settings.

The objectives selected have held up in examination after examination as appropriate for almost all types of teachers (vocational, special education, general education), in all grade levels and all subject areas. Although some redefinition of terms has been used (classroom management takes on new meanings when applied to developmentally delayed or behaviorally disordered students), and the relative importance of the different objectives has varied from one group of teachers to another, the fundamental structure of the objectives remains a cornerstone of the program.

Disadvantages

The research was, and probably still is, fragmentary with respect to characteristics of excellent teachers. What research there is has tended to concentrate on the characteristics which relate to instruction of younger students who are having, or would be expected to have, difficulty in mastering basic skills. Whether or not the identified characteristics will generalize satisfactorily to above average students, older students, and instruction related to higher cognitive levels or affective goals, remains to be seen. The Tennessee system found it necessary and desirable to go beyond the research in structuring an evaluation which would yield high scores for indirect as well as direct instructional methods.

Involvement of teachers was also a mixed blessing. Some tended to confuse involvement in the process with decision-making authority, and were unhappy when their favorite objectives were not included in the final draft. It also became apparent that many lacked the conceptual analysis skills to design a coherent, theoretically sound structure on which to base evaluation. This skill, which probably underlies the best instructional design, appears to be untaught and unrewarded in much of the preparation for and practice of teaching.

Develop instruments/procedures for data collection

In order to measure the competencies identified,it was necessary to identify likely sources of information about the teacher's performance in each area and to arrange instruments and procedures for collecting that information. It was thought, for example, that the information regarding teacher's ability to evaluate students might appropriately come from the teacher (via an interview or portfolio materials), the students (via interviews or questionnaires), the principal (via an interview or questionnaire), a written test, and possibly classroom observation. It was also thought that peer teachers probably would not know very much about a teacher's evaluation of students, and that the ability to identify this skill in the classroom observation would be somewhat limited.

For each area to be measured, as many possible sources of the required information as possible were identified. Existing instruments of all types (questionnaires, tests, observation schedules) were examined for their ability to provide information about the designated objectives. None was found to have an adequate level of content validity.

Accordingly, instruments were developed to measure the designated objectives by collecting data from each of the sources thought most likely to have the necessary information. The type of instrument used depended on the type of respondent (adults, students), the nature of the information required, and the cost/effectiveness of collecting the information in a reliable and objective way. It is important to note that the emphasis was on collecting reliable, objective information about each teacher. This information would then be scored and the results used to make judgments about the teacher's performance. Respondents were not asked to provide judgments directly.

For the first year of the program, seven instruments were developed. These and their current status are as follows:

Original instrument	Current status
1. Multiple-choice test	same
2. Classroom observation form	same
3. Structured candidate interview	Dialogue
4. Structured candidate portfolio	P. D. & L.
5. Structured principal interview	Prin. Questionnaire
6. Student questionnaire (primary, elementary and secondary forms)	same
7. Peer teacher questionnaire	(eliminated)

Additional information about the instruments and their relative strengths is available in McLarty (1987a).

The instruments were pilot and field tested in spring of 1983 and all except the primary version of the student questionnaire were found acceptable. The primary student questionnaire, designed for students in Kindergarten through grade 2, was not sufficiently reliable and was later revised, retested, and finally accepted for inclusion in the evaluation system.

All of the Career Ladder instruments were considered "secure" during the first year. Candidates were not provided access to the items on the Career Ladder test, the observation forms, or even the questionnaires used with their students, peers, and principals. There were two reasons for this approach: (1) to have the evaluation results be a reflection of typical (rather than maximal) teacher behavior, and (2) to have a score based on a limited sample of items generalize to a larger domain of teacher behaviors. If teachers knew the exact nature of the evaluation instruments, they might modify their behavior (or that of their students) in order to achieve a higher score which was not typical of their behavior. Teachers were, of course, provided with general descriptions of the

instruments and a complete descriptions of the competencies and indicators being assessed.

The security of the instruments did not make it through the first year. During the first year, the Tennessee Education Association released the secure rating scales for the Interview and Portfolio in order to "help teachers prepare better". This resulted in the Interview being dropped from the evaluation, and in a reduction of the weight assigned to some of the Portfolio items. In the second and succeeding years, all instruments except for the written test have been open, with copies provided to all candidates.

Advantages

The major advantage of the multiple data source methodology was to circumvent the weaknesses of each of the individual data sources. It seemed obvious that each of the sources considered had defects. The principal might have a personal grudge against a particular candidate; the evaluator might observe on an atypical day; the candidate might perform well on a written test but be unable to apply the knowledge in the practice of teaching. Combining information from various sources seemed the likeliest way of achieving a valid system which was fair to all teachers.

The design of new instruments was also an advantage in that it allowed each instrument to be keyed to the established objectives. The Tennessee observation form may well be the first criterion-referenced observation system to have been developed. This effort assured a system which was content valid.

Disadvantages

The most obvious disadvantages of a multiple data source system are the difficulty and cost. The start-up costs of developing new instruments are substantial. Pilot testing the instruments individually and then field testing them as a set required not only extensive staff and consultant commitment, but experimental subjects (randomly selected teachers) and field test evaluators. There is also the danger that a newly developed instrument will not function adequately and that one will be faced with a choice between using a less than ideal instrument and failing to meet timelines. In the case of the Tennessee Career Ladder, the primary form of the student questionnaire was delayed, redesigned, retested, and then implemented later than the remaining instruments.

Making the instruments secure was clearly a mistake. Although the intent (to measure typical rather than maximal behavior, and to have a small number of items generalize to

a larger domain of important behaviors) was appropriate, in retrospect it appears obvious that maintaining the security of the instruments when they were in the hands of more than a hundred evaluators travelling all over the state was unlikely to have been possible. The need to have a more open system also creates a need to have items which are themselves the desired behaviors (since they cannot be assumed to generalize to some larger domain of behaviors) and to recognize that maximal rather than typical behavior is probably being assessed.

Establish the scoring system

The method for scoring a multiple-choice test is usually a foregone conclusion: add up the items answered correctly, one point per item. Multiple data source systems offer no such easy solutions. In the Tennessee system there are thousands of individual pieces of information about every candidate (and this number varies by candidate because of varying numbers of students and observation scans). The data are multilevel (scans within observations within visits confounded with evaluator), and vary in raw score scale, distributional characteristics (most are skewed: questionnaires, negatively; observation scan data, positively), and reliability.

The challenge was to develop a scoring methodology which would appropriately combine these varying bits of information into a meaningful whole. After considerable discussion, it was decided to aggregate the information from each instrument over occasions, individuals, and (with some exceptions for the observation form) items addressing the same objective (domain and indicator). The scores so generated were placed on an integer stanine scale, weighted according to perceived importance and adequacy, and combined by domain.

The stanine methodology allowed a uniform, and, it was hoped, comprehensible scale, for the 80 individual scores generated for each candidate. It also adjusted for distributional characteristics of the scores (based on field test data, since standards had to be set in advance of actual evaluations). Reverse scored items were "righted" in the transformation so that "9" was always the highest score.

Despite the best of efforts and intentions, the stanine-based system did not work. The conversion to a stanine scale in scoring the principal questionnaire data resulted in some scores which were unanticipated by the principals and repugnant to them. Because of the wording of the items and the five-point response scale, and because of the extremely negatively skewed distributions encountered on the principal questionnaire field test data, principals who

marked items on the scale assuming a particular outcome for the teacher often encountered a less positive outcome and an angry teacher. (For more information on this situation, see McLarty, 1987b).

After the first year, the scaling methodology was changed to use a linear transformation of z-scores rather than the stanine approach. Each score was converted to a z-score scale (reversed if necessary), a constant added (for cosmetic purposes), and then weighted and summed. This continued to allow for the diverse scales and distributions but avoided the problems of trying to map five scale points onto nine. Once total domain scores were calculated, these were equated to the previous year and then translated to a 200-800 reporting scale.

For the current year, the z-score approach is being continued, but the means, standard deviations, and constants used are the same as for last year since no changes have been made in the instrumentation. This avoids the appearance of any kind of quota system and the need for any additional equating of results across years, since the same translation tables can continue to be used.

Advantages

The main advantage of the scoring system used by the Career Ladder is that the large number of scores provided to candidates on a uniform scale is diagnostic. A candidate can identify areas of strength and weakness by examining the individual scores, and, regardless of the Career Ladder level attained, can work to improve any weak areas. Since each score is associated both with a particular objective (domain and indicator) and with a single instrument, identification of the individual areas of need is fairly straightforeward.

In addition, the teacher is provided with each score on both the original (raw score) and the transformed (z-score) scale. This facilitates interpretation of the score with respect to both the instrument itself and the performance of other teachers. Additional information on score interpretation is available in McLarty (1987b).

Disadvantages

The complexity of the scoring system is its major disadvantage. The difficulty of generating the correct scores for each candidate (using the correct mean, standard deviation, constant, weight, and translation table) is exceeded only by the difficulty of explaining the scores once they are generated. First estimates of the time, staff, and equipment required were so far below the mark that a single mini-computer run to aggregate the 1983-84

observation data required more than 23 hours of CPU time, and the staff worked several 36 hour shifts to get the data processed.

The processing problems have mostly been resolved, but the complexity problems remain. Candidates continue to want to know exactly what they have to do to earn the desired status, and to feel that the data transformations are preventing them from fully understanding what is required. What they appear to want is a fully objective (quantitative) criterion-referenced evaluation for assessing merit in teaching. The current system is criterion-referenced in content, norm-referenced in scoring procedures, and has a fixed standard.

Set standards and cutoff scores

Once the instruments were developed and field tested, it was necessary to set standards for the evaluation system. Two standards were needed, one for Level II (the lower of the two merit levels), and one for Level III (the higher level). It was required that the standards not be based on quotas (i.e. not be norm-referenced), and that they be determined and announced in advance of any actual evaluations.

Unfortunately, the complexity of the data made standard-setting in the ways most commonly used for multiple-choice tests impractical. Thousands of bits of information were available regarding each teacher's performance, many of them replicated (over occasions as with the observation, or over individuals as with the students). No clear criterion was available, either, although each of the three evaluators was asked to provide a direct judgment of each candidate and the average of these judgments was used as an index of the candidate's general level of performance for standard setting purposes.

Distributions of scores for the field test candidates on the 80 scores originally created and on the five Domain scores were inspected and related to the average of the three evaluator judgments. Standards were then selected for Level II and Level III for each of the five domains. Based on the Panel's recommendation, it was decided to require performance in each of the five domains at or exceeding the Level II cutoff score to qualify for Level II. For Level III, the candidate would be required to score at or above the Level III cutoff score in four of the five domains and at or above the Level II cutoff score in the remaining domain. This decision was based on the recognition that candidates in the field test had not performed consistently well across the domains.

It was also decided that because none of the objective instruments could be expected to provide for all of the possible circumstances in which a teacher could demonstrate merit (or the lack thereof), the direct evaluator judgment would be included as an additional score. In order to provide as bias-free evaluation as possible, the three evaluators were to reach a consensus judgment about the candidate, and this judgment was to receive a weight in the decision in the same way as would any of the other scores. However, since the evaluators were providing judgments rather than information, the consensus judgments were made directly on the nine point decision scale rather than being transformed statistically to the stanine distribution.

Much of the discussion in the original standard-setting meetings focused on the anticipated increase in scores on the actual evaluation over the scores observed in the field test. Because of the extent to which score increases were expected to occur, and because the system was supposed to be "stringent", the standards selected were such that no field-test candidate would have qualified for Level III status. A more complete discussion of the original standard-setting effort can be found in McLarty (1986).

After the first year of implementation, when scaling was changed from a stanine approach to a z-score one, the linearly transformed z-scores were weighted and summed and the results expressed on a 200-800 scale. That reporting scale was the result of a common items equating process which fixed five scale points: the minimum (200), the qualifying score to be considered for Career Ladder II (450), the cutoff score for Level II (500), and cutoff score for Level III (700), and the maximum (800). All of the points were equated to the first year's equivalent scores by domain in order to maintain standards despite some changes in items and instruments.

In addition to changing the scaling process and reporting scales, the second year of implementation the Career Ladder moved to the use of a single cut-off score, based on a weighted combination of the five domain scores, as the criterion. In order to compensate for the previous domain-by-domain qualification rules, minimum qualifying scores were set for each domain. It was necessary to meet these scores in order to be considered for the higher levels of the Career Ladder.

In the third year of implementation, the same scoring process is being used as for the second year except that no new equating is being done. Instead, no changes have been made in items or instruments, so the previous year's means, standard deviations, constants, and equating tables can be used.

Advantages

The biggest advantage of the current standards is that they exist and they appear to be working. When the legislature dictated that the Career Ladder would not have quotas but rather that the standards should be "stringent", it is unlikely that they realized the difficulty of the task. Working with new, highly quantitative instruments and setting standards in advance of any actual evaluations virtually guaranteed that the field test data would not be an accurate representation of actual teacher performance. This made the identification of the instruments' stringency levels nearly impossible. The data collection approach to measurement (as opposed to a judgment-based approach) also made the use of item/content-based standard setting methods impractical. No one knew what frequency of task-description or monitoring should be required for a Level III teacher, or whether that standard should vary (for example by grade level and subject area). Eventually, the Standard-setting panel simply had to make the necessary decisions and hope for the best.

The Tennessee Career Ladder is a very visible system. As a result, it is validated (or invalidated) every time candidate results are announced. The first year, a number of teachers who probably should have qualified missed doing so because of problems with scoring in the Planning domain. Those who did qualify, however, were generally seen as deserving. The second year the problem with the Planning domain was solved by using a weighted total score. However this (and several other) changes made the system somewhat easier than previously despite the equating of the objective portion of the data and a number of teachers not seen as deserving apparently qualified.. In both years, however, the vast majority of the teachers who have qualified for the upper levels of the Career Ladder have been seen as deserving of that status. Considering the nature of the timelines and information available on which to base standards, and the various requirements placed on standard-setters, this is a remarkable achievement.

Disadvantages

There are two main disadvantages of the current Career Ladder standards. The first, and probably most important, disadvantage is that the current standards do not have intrinsic meaning. Because of the complexity of the data itself, the weighted sum of the various scores does not have meaning other than with respect to the standards. Several attempts to simplify the scoring and establish more meaningful standards have been made, but these have not been successful. The main problems encountered have related to the severe skews in the data.

For example, one standard-setting panel which tried to establish direct relationships between the individual raw score scales for the seventy scores in the system and the final decision scale gave up when they realized that on some questionnaire data seventy percent of the teachers had achieved perfect raw scores. The panel was unable to accept scaling the "5" on a scale of 1-5 as an "800" on the 200-800 reporting scale because they felt that if that many candidates had achieved perfection, it was likely that to translate the score to a hundred points higher than the Level III cutoff score would be too high. However, since it was not possible for a teacher to score any higher on the raw score scale, they also could not justify scaling that score any lower then a perfect score on the 200-800 scale.

The panel finally concluded that the current practice of weighting and combining scores before translating them to the final reporting scale was the best available method. Although it does not solve the problem of floor and ceiling effects for individual scores, the method does subsume them in the more comprehensive domain and total scores.

The second disadvantage is that they standards are still being linked to three year old field test data (which wasn't all that good in the first place). Even the original scoring problems with the Planning domain have been faithfully reproduced in the equating. The standards need to be reevaluated and reset. With the additional information and experience gained during the last three years, it should be possible to establish new and more meaningful standards for the Career Ladder evaluation system.

Evaluation and validation of the system

As mentioned above, the Tennessee Career Ladder evaluation system is validated constantly on a informal basis. In addition, data have been collected from candidates, evaluators, and others regarding the individual instruments, items, procedures, and personnel. Data analysis has been used to make a number of improvements in the system on an ongoing basis. In addition, a more formal validation study has been begun. This study, which is mandated in the law, will probably take several years to complete.

Probably the most common validation question to be asked is whether the "right" teachers have been identified as meritorious. While this is an important question, it is neither the most important question nor one which can be given a unambiguous answer. The lack of a criterion by which to judge whether the "right" teachers were chosen

accounts for the lack of an unambiguous answer to the question. Short of finding a clearly defensible means of evaluating the relevant student achievements attributable to each individual teacher, validation studies of the Tennessee Career Ladder evaluation system will probably have to be based on multiple criteria. This presents some unique problems in Tennessee because the multiple criteria have already been included in the evaluation itself.

Beyond the "correct teacher" question, however, is the more important question of whether in fact Career Ladder programs do act to improve education, directly or indirectly. In Tennessee, it is not likely to be possible to provide definitive information on this question. This is because the effects of a Career Ladder program on education generally, particularly the indirect effects associated with recruiting and retaining the best teachers and motivating all teachers to perform at their best, are likely to depend on the teachers' perception of the permanency of the program. Only when the program is an integral part of the teacher's (or would-be teacher's) decision process, is it likely to have the desired effect.

In Tennessee, the Career Ladder program has not established the degree of stability necessary to become a part of the teacher's career plan. The object of constant assaults by the Tennessee Education Association (TEA), Career Ladder has had to expend considerable resources in time, staff, and effort just to continue to exist. Each year, at least one bill has been introduced in the Legislature to abolish the program. Constant rhetoric by the TEA assures legislators and teachers that the program is not working, and that "most" teachers are against it.

With the election new, TEA-backed administration, the program has been further devalued. Accusations about excessive costs and bureaucracy have been voiced and published, creating further doubts in the minds of teachers. Under these circumstances, the program is not likely to have the indirect effects it was intended to have, regardless of the quality of the evaluation system itself.

Summary

The Tennessee Career Ladder program is a complex, multiple-data-source system for evaluating teachers. Since its implementation almost three years ago, the program has been strengthened and improved. Most of the original problems have been solved, and those remaining appear to be solvable. Although a definitive validation study has yet to be completed, the evaluation system appears to be identifying meritorious teachers accurately in most cases.

The major disadvantage of the system is its level of sophistication. The complexity of good teaching, the high-stakes nature of the award (up to $35,000 per candidate over a five year period), and the lack of any quota system (no external controls on the total cost of the awards), led the developers of the program to provide the most accurate, fair, and carefully equated program possible. Such a program is necessarily costly and complex. However, the additional accuracy and fairness possible with the multiple-data-source approach is essential where Career Ladder decisions which have substantial financial and personal impact must be made.

Under the new administration, it is likely that the Tennessee Career Ladder program will evolve in new ways. In evaluating the changes in the program and their effects, it will be especially important to identify changes in the goals of the program (for example toward support of direct rather than indirect effects on education), and in the resources directed toward their accomplishment.

References

Furtwengler, C., Malo, G., McLarty, J. R., and Strouss, S. (April, 1986). Multiple data sources in teacher evaluation. Paper presented at the National Council on Measurement in Education annual meeting, San Francisco, CA. (ERIC Document Reproduction Service No. TM 860 536).

McLarty, J. R. (April, 1987a). Career Ladder instrumentation: The Tennessee Experience. Symposium presentation at the American Educational Research Association annual meeting, Washington, D. C.

McLarty, J. R. (April, 1987b). Perspectives on the summative evaluation of teachers: the reporting perspective. Symposium presentation at the American Educational Research Association and National Council on Measurement in Education annual meetings, Washington, D. C.

McLarty, J. R., Furtwengler, C., & Malo, G. (April, 1985). Using multiple data sources in teacher evaluation. Paper presented at the National Council on Measurement in Education annual meeting, Chicago, IL.

FIGURE I

STEPS IN DEVELOPING CAREER LADDER EVALUATION SYSTEMS

1. FORMULATE CRITERIA/COMPETENCIES TO BE ASSESSED

   SOURCES: 1. RESEARCH ON EFFECTIVE PERFORMANCE
   2. CONSENSUS OF THOSE TO BE EVALUATED

2. VERIFY CRITERIA/COMPETENCIES

3. DETERMINE VALID SOURCES OF DATA

4. DEVELOP INSTRUMENTS/PROCEDURES FOR DATA COLLECTION

5. PILOT INSTRUMENTS/PROCEDURES FOR DATA COLLECTION

6. ESTABLISH SCORING SYSTEM

7. DEVELOP EVALUATOR TRAINING MATERIALS AND PROCEDURES

8. SELECT AND TRAIN FIELD TEST EVALUATIONS

9. CONDUCT FIELD TEST OF ENTIRE EVALUATION SYSTEM

   PURPOSE: 1. DETERMINE VALIDITY AND RELIABILITY OF THE SYSTEM
   2. COLLECT DATA TO BE USED IN SETTING STANDARDS

10. REVISE SYSTEM BASED ON FIELD TEST DATA

11. SET STANDARDS AND CUTOFF SCORES

12. DEVELOP ORIENTATION MANUAL FOR CANDIDATES

13. CONDUCT ORIENTATION FOR CANDIDATES

14. RETRAIN FIELD TEST EVALUATORS/TRAIN NEW EVALUATORS

Appendix A

Data Sources for the Tennessee Career Ladder

Competencies and Indicators

for

1984-85

&

1985-86

1984-5

Data Sources for

COMPETENCIES AND INDICATORS

	Classroom Observation	Portfolio	Applicant Interview	Principal Interview	Peer Questionnaire	Student Questionnaire	Written Test
I. Prepares for instruction effectively.							X
A. Establishes appropriate instructional goals and related objectives consistent with the curriculum.		X	X	X			
B. Prepares instructional plans and materials incorporating principles of effective instruction.	X	X	X	X			
C. Creates, selects or modifies instructional plans and materials to accommodate learner instructional levels.	X	X	X	X			
II. Uses teaching strategies and procedures appropriate to the content, objectives, and learners.							X
A. Provides a clear description of the learning task and its content.	X	X				X	
B. Monitors learner understanding and reteaches as necessary.	X	X	X			X	
C. Provides learners appropriate practice and review.	X	X	X			X	
D. Establishes and maintains learner involvement in the learning task.	X	X	X			X	
III. Uses evaluation to improve instruction.							X
A. Uses information about learner performance to improve the instructional process.		X	X	X			
B. Reports learner status and progress to learners and their parents.	X	X	X	X		X	
IV. Manages classroom activities effectively.							X
A. Establishes and maintains appropriate learner behavior.	X	X		X		X	
B. Establishes and maintains a classroom climate conducive to learning.	X	X	X	X		X	
C. Makes effective use of classroom resources (e.g. personnel, time, materials, facilities)	X		X	X		X	
V. Establishes and maintains a professional leadership role.							
A. Improves professional skills and knowledge.		X	X	X			
B. Takes a leadership role in improving education.		X	X	X	X		
C. (Screening only) Performs professional responsibilities efficiently.			X	X	X		
VI. (Screening only) Communicates effectively.							X
A. Writes clearly and correctly.	X					X	
B. Communicates oral information effectively.	X		X	X		X	
C. Reads professionally relevant literature/materials with comprehension.		X	X	X			

HOW WE OBTAIN OUR INFORMATION

ABOUT

COMPETENCIES AND INDICATORS

GENERAL EDUCATION

	CLASSROOM OBSERVATION	DIALOGUE	PEER QUESTIONNAIRE	PRINCIPAL QUESTIONNAIRE	STUDENT QUESTIONNAIRE	PROFESSIONAL DEVELOPMENT AND LEADERSHIP SURVEY	WRITTEN TEST
I. Prepares for instruction effectively.							X
A. Establishes appropriate instructional goals and related objectives consistent with the curriculum.		X		X			
B. Prepares instructional plans and materials incorporating principles of effective instruction.	X	X		X			
C. Creates, selects or modifies instructional plans and materials to accommodate learner instructional levels.	X	X		X			
II. Uses teaching strategies and procedures appropriate to the content, objectives, and learners.							X
A. Provides a clear description of the learning task and its content.	X	X		X	X		
B. Monitors learner understanding and reteaches as necessary.	X	X		X	X		
C. Provides learners appropriate practice and review.	X	X		X	X		
D. Establishes and maintains learner involvement in the learning task.	X	X		X	X		
III. Uses evaluation to improve instruction.							X
A. Uses information about learner performance to improve the instructional process.		X		X			
B. Reports learner status and progress to learners and their parents.	X	X		X	X		
C. Improves learner performance (Field Test Only)							
IV. Manages classroom activities effectively.							X
A. Establishes and maintains appropriate learner behavior.	X			X	X		
B. Establishes and maintains a classroom climate conducive to learning.	X	X		X	X		
C. Makes effective use of classroom resources (e.g. personnel, time, materials, facilities)	X			X	X		
V. Establishes and maintains a professional leadership role.							
A. Improves professional skills and knowledge.				X		X	
B. Takes a leadership role in improving education.			X	X		X	
C. (Screening only) Performs professional responsibilities efficiently.							
VI. (Screening only) Communicates effectively.							X
A. Writes clearly and correctly.							X
B. Reads professionally relevant literature/materials with comprehension.							X